

Leveraging Large Language Models for Recommender Systems

A snapshot of the state-of-the-art

Dietmar Jannach, University of Klagenfurt, Austria

dietmar.jannach@aau.at

Presented at the “Workshop on Generative AI for Recommender Systems and Personalization” ,
co-located with KDD '24

Barcelona, 2024

Recommender Systems

- A pervasive part of our daily online user experience
- One of the most widely used applications of machine learning

You may also like



Jack & Jones
JAMIE - Polo shirt - orange
£21.00
Free delivery & returns

Related hotels...



Hotel 41
★★★★★ 1,170 Reviews
London, England

Show Prices



Jobs you may be interested in ^{Beta} [Email Alerts](#) | [See More »](#)

	Technical Sales Manager - Europe Thermal Transfer Products - Home office	×
	Senior Program Manager (f/m) Johnson Controls - Germany-NW-Burscheid	×




You may also like

 ★★★★☆ (109)	 ★★★★★ (53)	 ★★★★☆ (33)
--	---	--

Read Commented **Recommended**

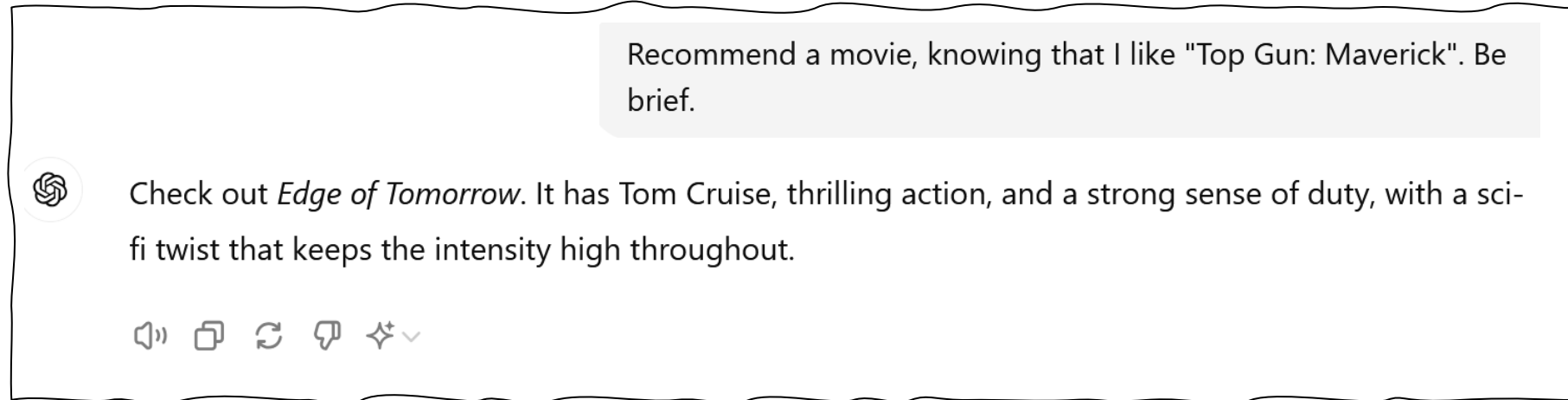
	Germany Just Rejected The Idea That The European Bailout Fund Would Buy Spanish Debt	×
	There Is Almost No Gold In The Olympic Gold Medal	×

Groups You May Like [More »](#)

	Advances in Preference Handling Join
	FP7 Information and Communication Technologies (ICT) Join
	The Blakemore Foundation Join

Large Language Models

- A **future** pervasive part of our daily online user experience?



- Meaningful recommendations without model training ✓
- Explanation and persuasion, customizable ✓
- Natural language interaction ✓

Large Language Models and RecSys

- Significant developments since ChatGPT / 2022
- Huge interest in leveraging LLMs and AI assistants
- Feels like every second RecSys paper is involving LLMs

The screenshot shows a Google Scholar search for "llm recommender system". The search results page includes a sidebar on the left with filters and a main content area with search results. Red circles highlight specific elements: the search bar, the result count, the "Beliebige Zeit" filter, and the "Zitate einschließen" checkbox.

Google Scholar

llm recommender system

Artikel

Ungefähr 10 500 Ergebnisse (0,05 Sek.)

Beliebige Zeit

Seit 2024

Seit 2023

Seit 2020

Zeitraum wählen...

Nach Relevanz sortieren

Nach Datum sortieren

Beliebige Sprache

Seiten auf Deutsch

Alle Typen

Übersichtsarbeiten

Patente einschließen

Zitate einschließen

OpenP5: An Open-Source Platform for Developing, Training, and Evaluating LLM-based Recommender Systems

S Xu, W Hua, Y Zhang - Proceedings of the 47th International ACM SIGIR ..., 2024 - dl.acm.org

... models into **recommender systems**. ... of **LLM**, how to adapt **LLM**, and the **recommendation** tasks. With the powerful ability, **LLM** can participate in several components of **recommender** ...

☆ Speichern Zitieren Zitiert von: 2 Ähnliche Artikel Alle 2 Versionen

LLM-generated Explanations for Recommender Systems

S Lubos, TNT Tran, A Felfernjg... - Adjunct Proceedings of ..., 2024 - dl.acm.org

... that **LLM**-generated explanations are highly appreciated by users as they help in the evaluation of **recommended** items. Furthermore, we discuss which characteristics of the **LLM**-based ...

☆ Speichern Zitieren Zitiert von: 1 Ähnliche Artikel Alle 2 Versionen

Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System

S Kim, H Kang, S Choi, D Kim, M Yang... - arXiv preprint arXiv ..., 2024 - arxiv.org

... as follows: • We present an **LLM**-based **recommender system**, called A-LLMRec, that ...

recommender system. • A-LLMRec requires the fine-tuning of neither the CF-RecSys nor the **LLM**, ...

☆ Speichern Zitieren Zitiert von: 1 Ähnliche Artikel Alle 2 Versionen

Outline

- Part I:
 - A survey of surveys
- Part II:
 - Overview of some own recent works
 - LLMs for Sequential Recommendation
 - Human Evaluation of ChatGPT recommendations
 - Beyond LLMs: Image Generation for Fashion Recommendation
- Slides:
 - <https://tinyurl.com/kdd24-genai>

Part I: Survey of surveys

Survey papers (including preprints)

- A Survey on Large Language Models for Recommendation (May 2023)
- How Can Recommender Systems Benefit from Large Language Models: A Survey (June 2023)
- Recommender Systems in the Era of Large Language Models (LLMs) (July 2023)
- Large Language Models for Generative Recommendation: A Survey and Visionary Discussion (Sep 2023)
- Foundation Models for Recommender Systems: A Survey and New Perspectives (Feb 2024)
- A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys) (Mar 2024)
- ...



A Survey on Large Language Models for Recommendation

Likang Wu^{1,2*}, **Zhi Zheng^{1,2*}**, **Zhaopeng Qiu^{2*}**, **Hao Wang^{1†}**, **Hongchao Gu¹**, **Tingjia Shen¹**, **Chuan Qin²**, **Chen Zhu²**, **Hengshu Zhu^{2†}**, **Qi Liu¹**, **Hui Xiong^{3†}**, **Enhong Chen^{1†}**

¹University of Science and Technology of China, ²Career Science Lab, BOSS Zhipin, ³Hong Kong University of Science and Technology (Guangzhou)

{wulk,zhengzhi97,hcgu,jts_stj}@mail.ustc.edu.cn,

{zhpengqiu,chuanqin0426,zc3930155,zhuhengshu}@gmail.com,

{wanghao3,qiliuql,cheneh}@ustc.edu.cn, xionghui@ust.hk



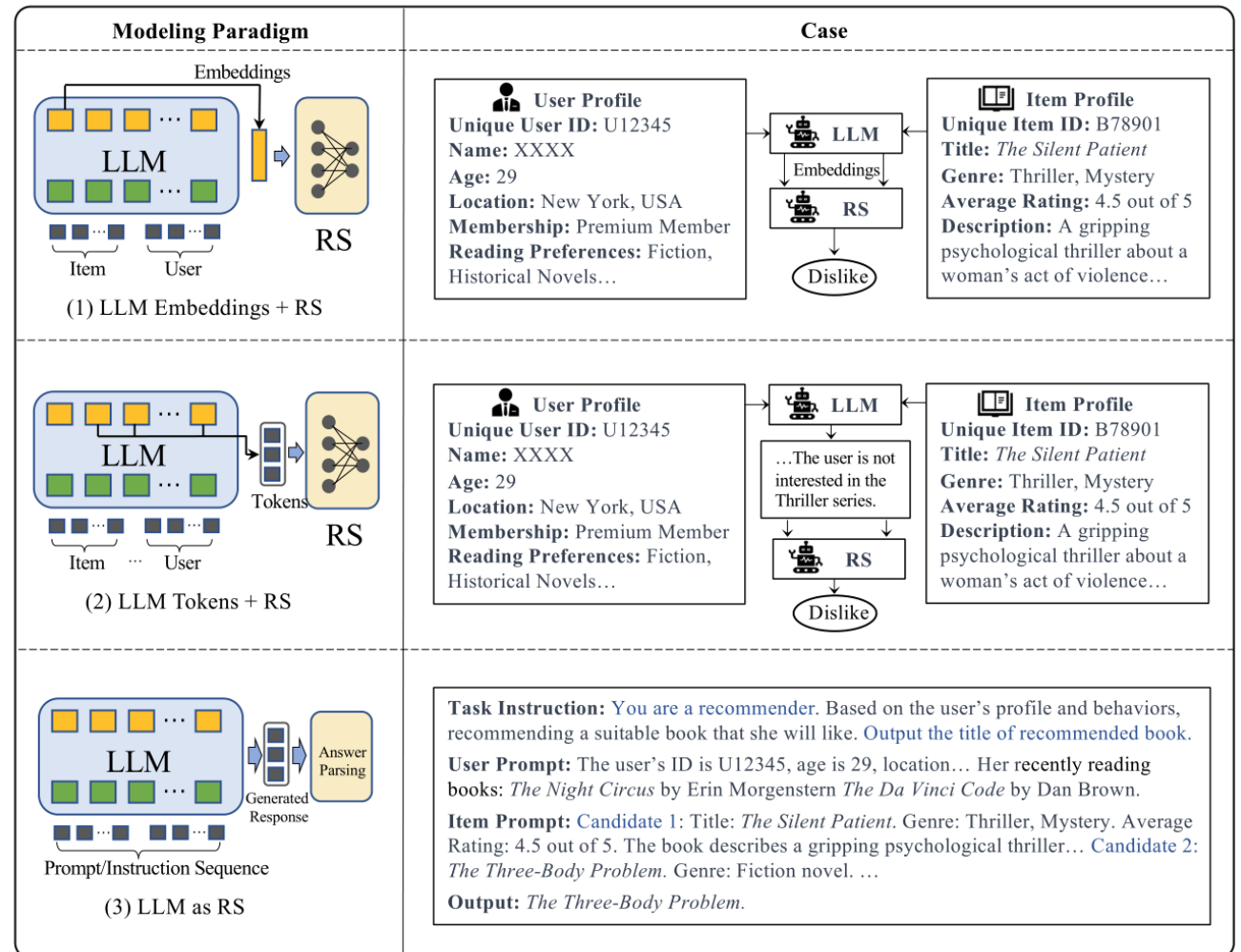
A Survey on Large Language Models for Recommendation

“The key advantage of incorporating LLMs into recommendation systems lies in their ability to extract high-quality representations of textual features and leverage the extensive external knowledge encoded within them”

- Distinguishes between **discriminative** LLMs and **generative** LLMs for recommendation (LLM4Rec)
 - **Discriminative LLM4Rec**: Mostly BERT-based models and approaches
 - **Generative LLM4Rec**: Models with better generative capabilities, e.g., the GPT series

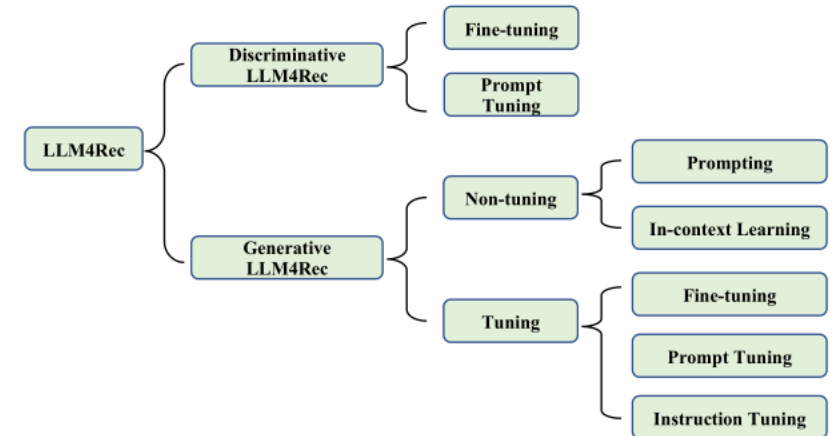
A Survey on Large Language Models for Recommendation

- Identifies three modeling paradigms:
 - LLM Embeddings + RS:** The LLM serves as a feature extractor for enhanced user and/or item embeddings
 - LLM Embeddings + Tokens:** Similar to the previous ones, but generates tokens
 - LLM as RS:** A direct approach, where the input is a profile and a task prompt, and the output is a recommendation



A Survey on Large Language Models for Recommendation

- Identifies different forms of tuning
 - **Non-tuning:** Leverage built-in capabilities of LLMs. Challenge is find good instructions/prompts and/or provide demonstration examples without re-training
 - **Tuning:**
 - **Fine-tuning:** Use a pre-trained language model and enhance it with a recommendation specific dataset
 - **Prompt tuning:** Specific tuning approaches to learn to generate preference assessments or items to recommend
 - **Instruction tuning:** Fine-tuning an LLM for different tasks with different instructions, e.g., sequential recommendation, explanation, rating prediction



How Can Recommender Systems Benefit from Large Language Models: A Survey

JIANGHAO LIN^{*}, Shanghai Jiao Tong University, China

XINYI DAI^{*}, Noah's Ark Lab, Huawei, China

YUNJIA XI, Shanghai Jiao Tong University, China

WEIWEN LIU and BO CHEN, Noah's Ark Lab, Huawei, China

HAO ZHANG and YONG LIU, Noah's Ark Lab, Huawei, Singapore

CHUHAN WU and XIANGYANG LI, Noah's Ark Lab, Huawei, China

CHENXU ZHU and HUIFENG GUO, Noah's Ark Lab, Huawei, China

YONG YU, Shanghai Jiao Tong University, China

RUIMING TANG[†], Noah's Ark Lab, Huawei, China

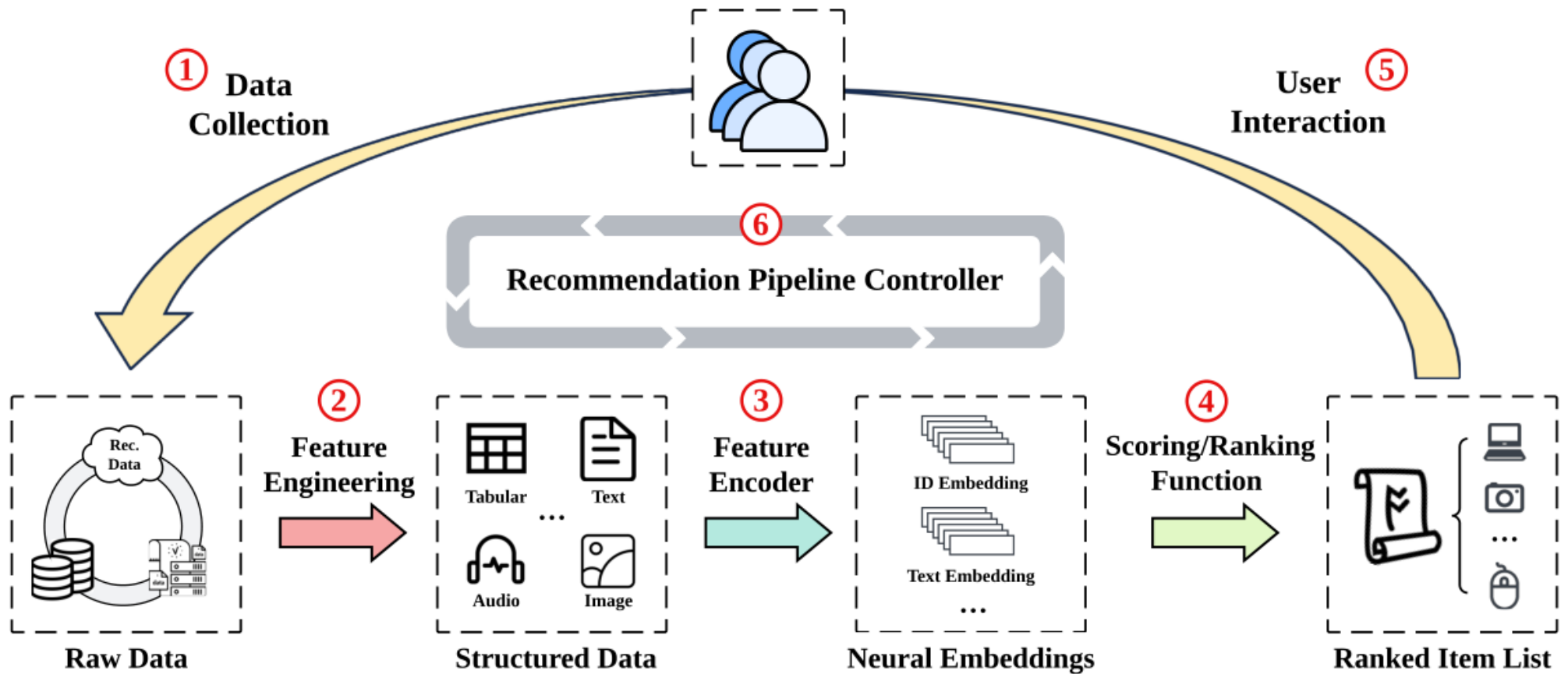
WEINAN ZHANG[†], Shanghai Jiao Tong University, China

How Can Recommender Systems Benefit from Large Language Models: A Survey

- A survey from the angle of the whole pipeline in industrial recommender systems
- Two perspectives, WHERE and HOW
 - **WHERE**: at which stages in the pipeline
 - **HOW**: how to adapt LLMs for RS
 - Freezing the parameters of the LLM or not
 - Involve a conventional recommendation model (CRM) or not

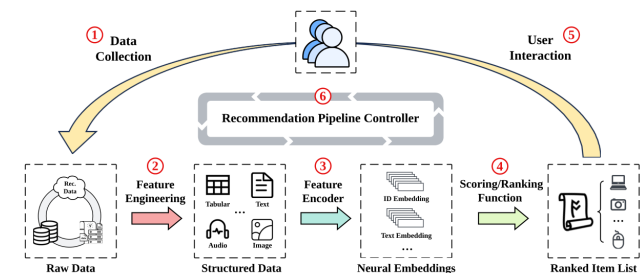
How Can Recommender Systems Benefit from Large Language Models: A Survey

- WHERE:



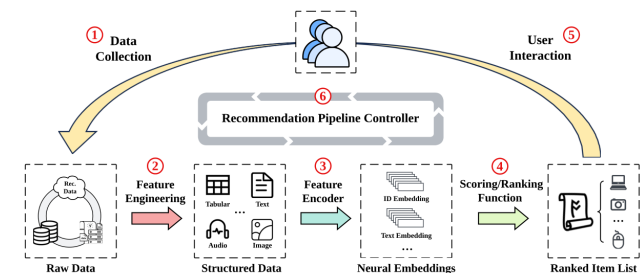
How Can Recommender Systems Benefit from Large Language Models: A Survey (WHERE)

- LLM for Feature Engineering
 - LLM takes original features and generates additional textual features for data augmentation
- LLM as a Feature Encoder
 - Representation Enhancement for better embeddings
 - Unified Cross-Domain Recommendation, supporting transfer learning and cross-domain recommendation



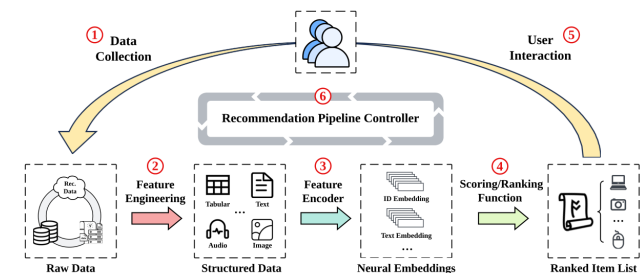
How Can Recommender Systems Benefit from Large Language Models: A Survey (WHERE)

- LLMs as Scoring/Ranking Function:
 - Item Scoring Task
 - Item Generation Task
 - Open-set (direct generation of items to recommend)
 - Closed-set (with a candidate list to rank)
 - Hybrid, with LLM used both for generation and ranking



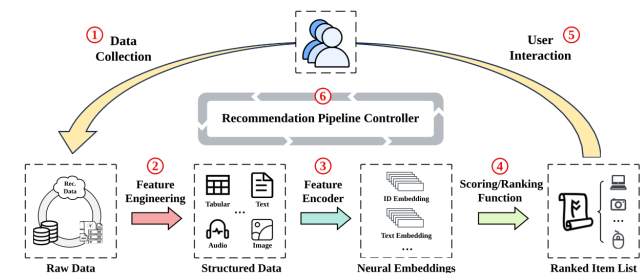
How Can Recommender Systems Benefit from Large Language Models: A Survey (WHERE)

- LLM for User Interaction in a conversational system
 - Task-oriented User Interaction
 - Assumes that the user has a clear intent
 - LLM for example used to understand the user intentions or for explanations
 - Open-ended User Interaction
 - Assumes only a vague user intent
 - Goal is to acquire interests and preferences
 - The LLM drives the interaction



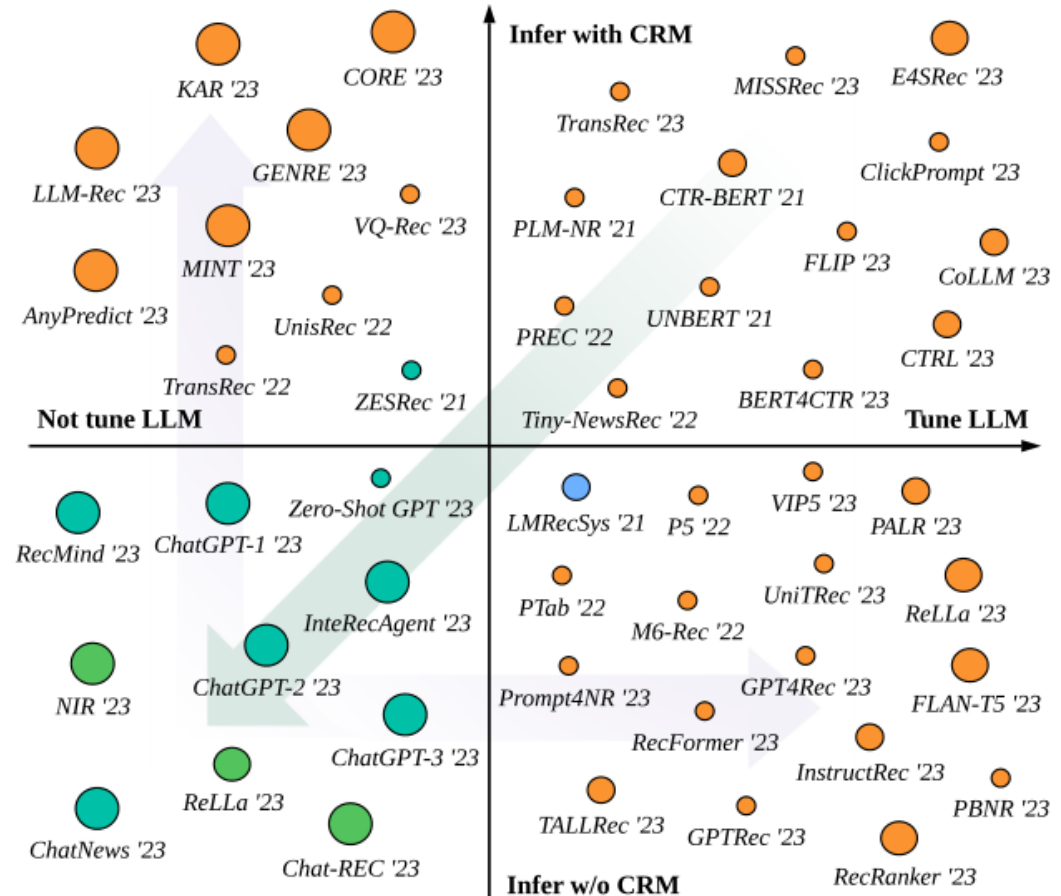
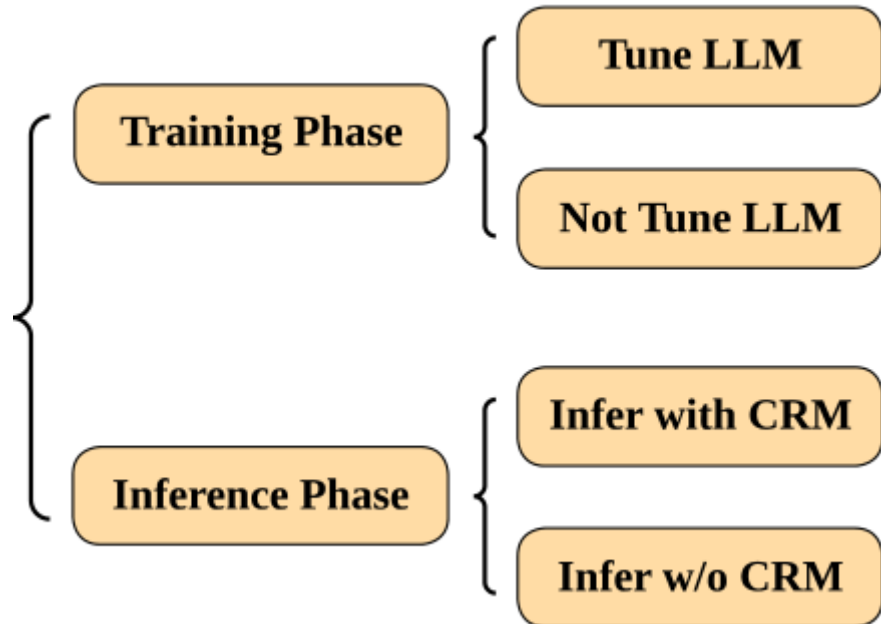
How Can Recommender Systems Benefit from Large Language Models: A Survey (WHERE)


- LLM as a Pipeline Controller
 - LLM not only one component of the RS, but could “*participate in the pipeline control of the system.*”
 - LLM is used to make decisions, e.g., in a conversational setting, decide if recommendations should be presented or more preferences should be acquired.



How Can Recommender Systems Benefit from Large Language Models: A Survey


- HOW:





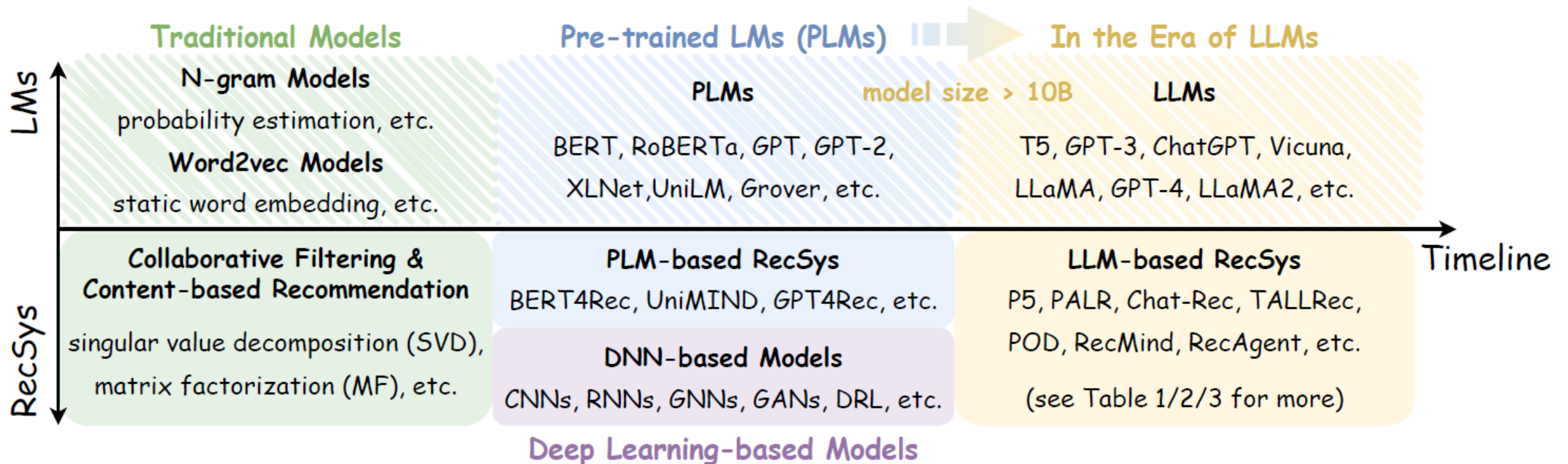
Recommender Systems in the Era of Large Language Models (LLMs)

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang,
Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li



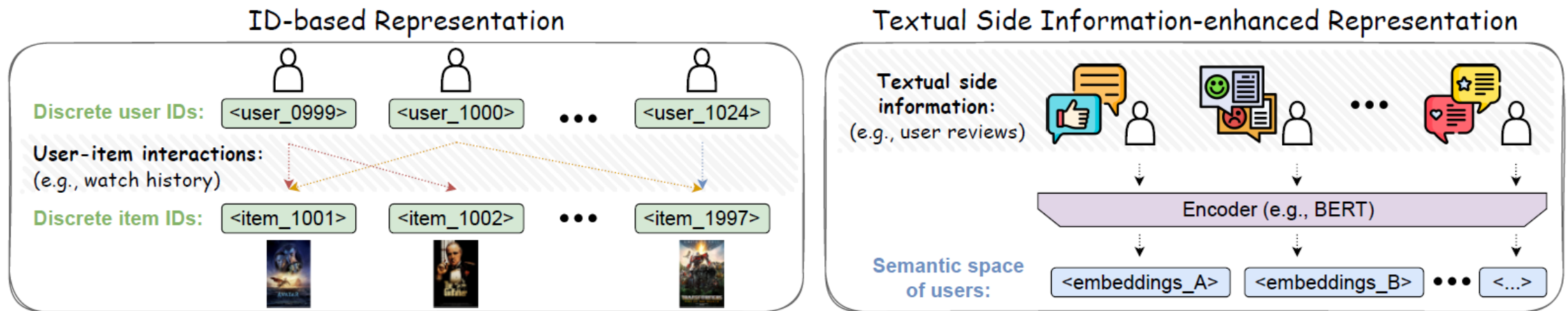
Recommender Systems in the Era of Large Language Models

- Shares categorization of first paper (fine-tuning, prompting, ..)
- But also provides some history



Recommender Systems in the Era of Large Language Models

- Alternative entity representations in LLMs





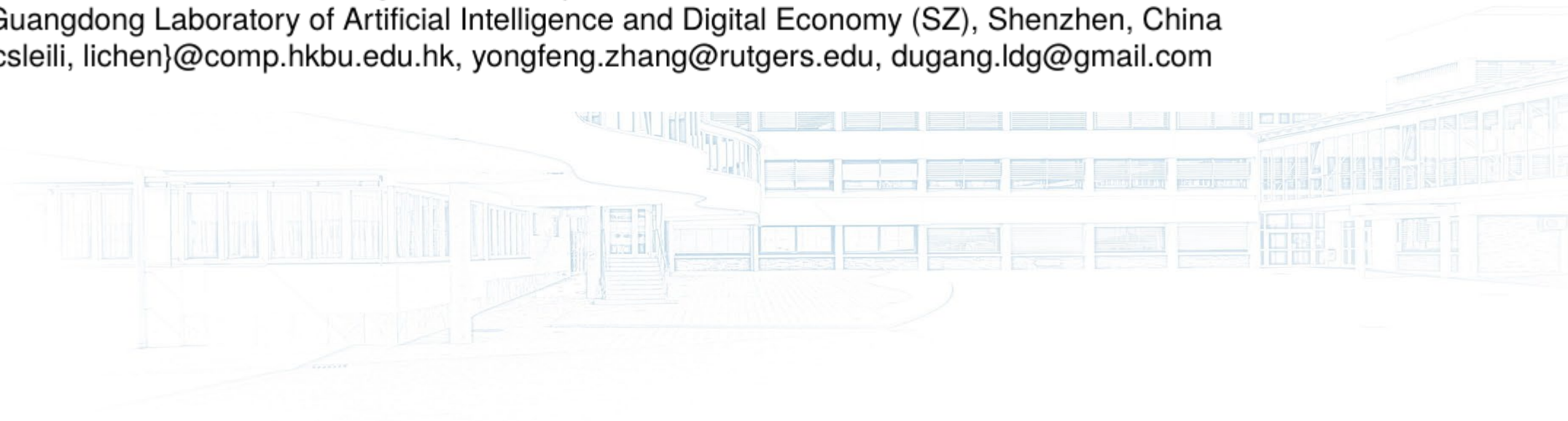
Large Language Models for Generative Recommendation: A Survey and Visionary Discussions

Lei Li¹, Yongfeng Zhang², Dugang Liu³, Li Chen¹

¹Hong Kong Baptist University, Hong Kong, China

²Rutgers University, New Brunswick, USA

³Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China
{csleili, lichen}@comp.hkbu.edu.hk, yongfeng.zhang@rutgers.edu, dugang.ldg@gmail.com



Large Language Models for Generative Recommendation: A Survey and ...

- Three main dimensions
 - **What** is generative recommendation?
 - **Why** we should advance to generative RS
 - How to implement LLM-based generative recommendations
- Provides a new taxonomy and a definition (**what**)

Definition 2 (Generative Recommendation) *A generative recommender system directly generates recommendations or recommendation-related content without the need to calculate each candidate's ranking score one by one.*

- **Why** using it:
 - To avoid the need for common two-stage ranking pipelines


Large Language Models for Generative Recommendation: A Survey and ...

- **How: ID Creation**

- Items must be identified somehow
 - Traditionally done by item IDs
- For LLMs, item descriptions could be IDs
 - But they can be long; and exact matches can be difficult
- Embeddings could be IDs as well
 - But they require much memory
- **Proposal here:** Use small number of tokens to represent items
 - Different approaches from the literature discussed

Large Language Models for Generative Recommendation: A Survey and ...

- **How:** Provides examples for prompt-based approaches for
 - rating prediction, top-n recommendation, sequential recommendation, explanation generation, review generation, review summarization
- **How:** Conversational recommendation
 - Either provide dialog snippets and ask LLM to respond; or directly interact with the LLM-based AI assistant
- How to evaluate?
 - Focus on common offline quality measures



Foundation Models for Recommender Systems: A Survey and New Perspectives

Chengkai Huang¹, **Tong Yu²**, **Kaige Xie³**, **Shuai Zhang⁴**, **Lina Yao^{1,5}** and **Julian McAuley⁶**

¹The University of New South Wales

²Adobe Research

³Georgia Institute of Technology

⁴ETH Zurich

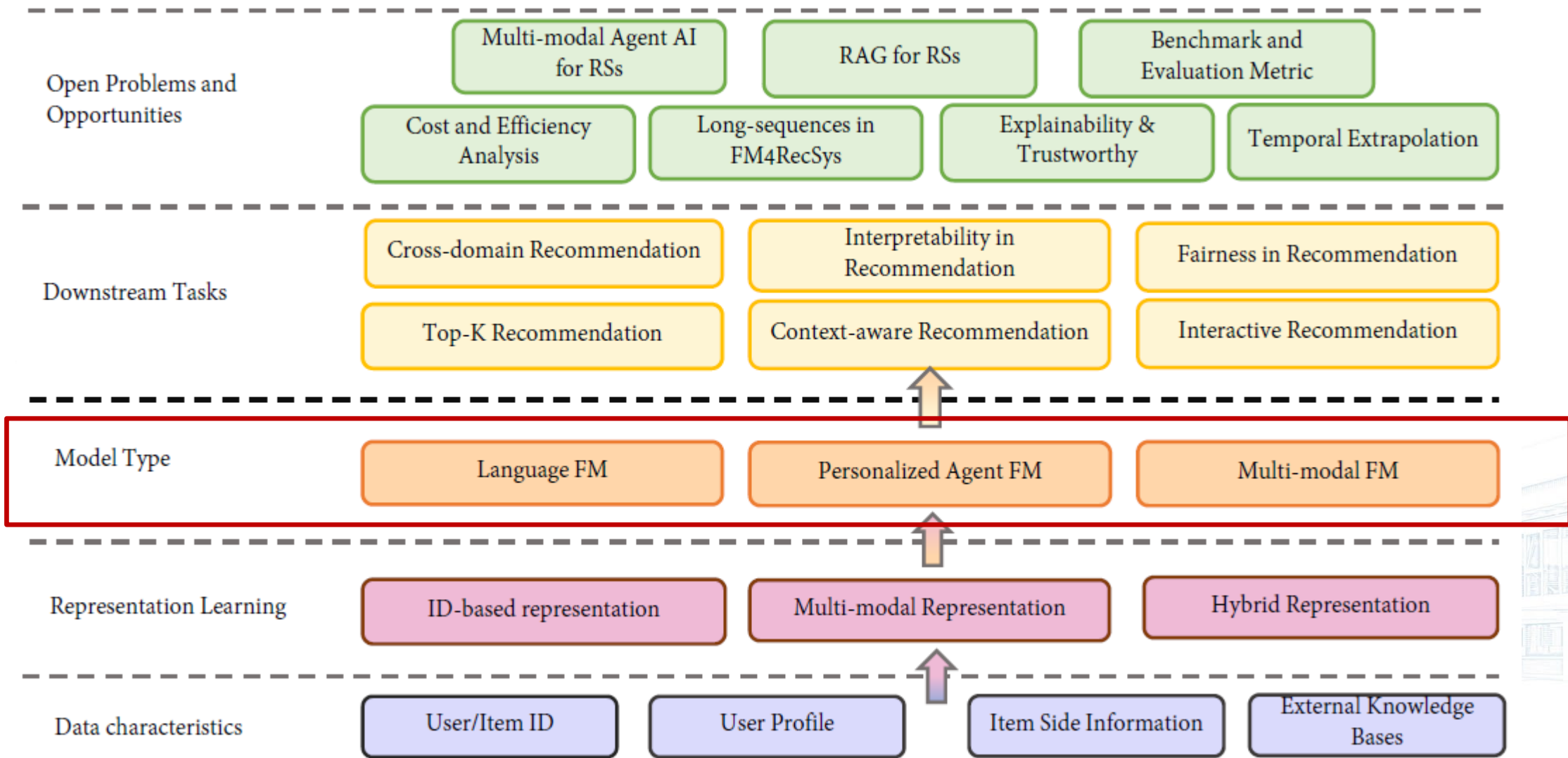
⁵CSIRO's Data61

⁶University of California San Diego

{chengkai.huang1, lina.yao}@unsw.edu.au, tyu@adobe.com, kaigexie@gatech.edu,
cheungshuai@outlook.com, lina.yao@data61.csiro.au, jmcauley@eng.ucsd.edu

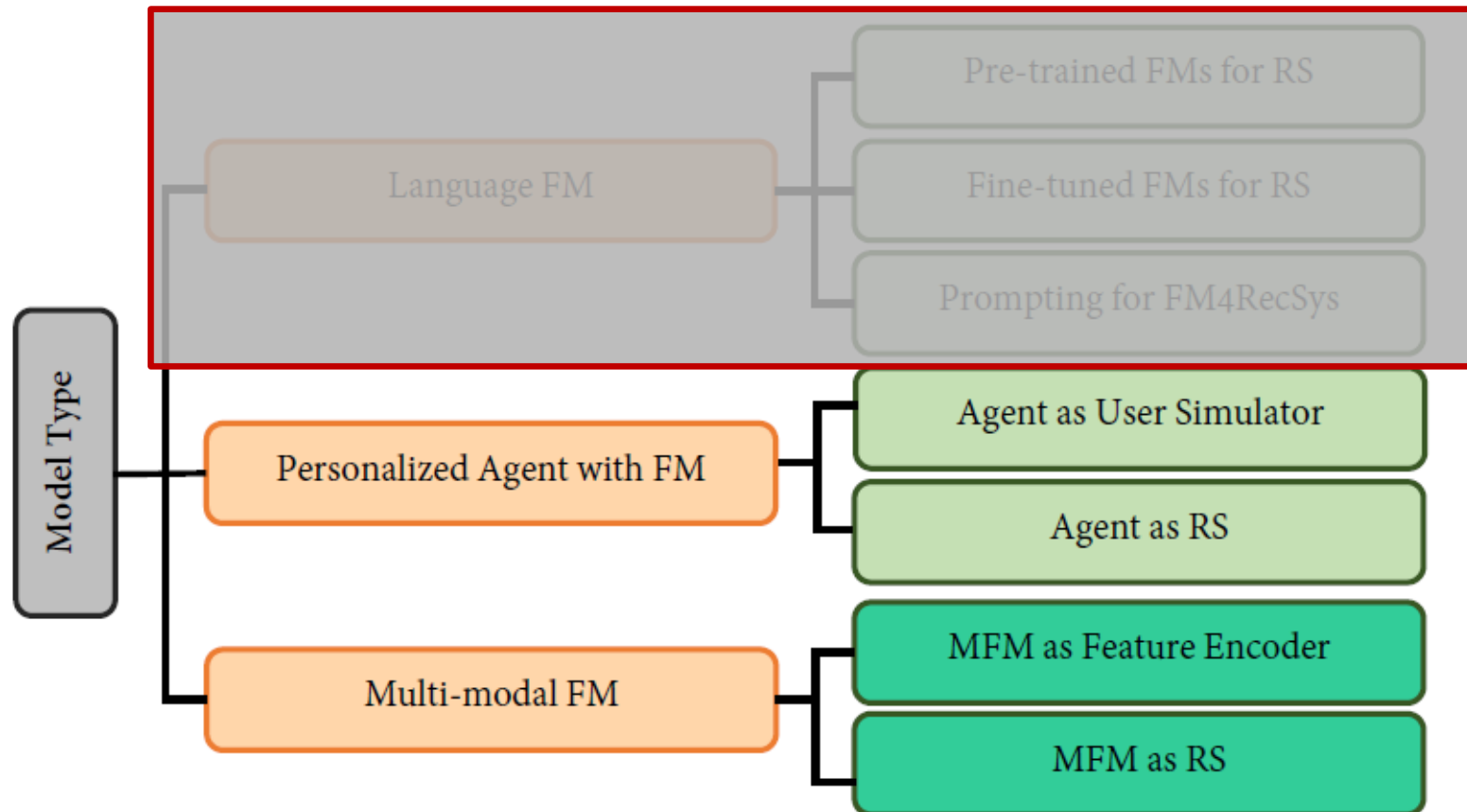
Foundation Models for Recommender Systems: A Survey and New Perspectives

- FM4RecSys
 - Survey intended to go beyond LLMs
 - Consider wider range of foundation models
 - multi-modal representation, hybrid representation
 - Provide a framework of using FMs in recommender systems
 - considering various downstream tasks
 - including open problems and opportunities



Foundation Models for Recommender Systems: A Survey and New Perspectives

- Different model types, with LLMs being predominant

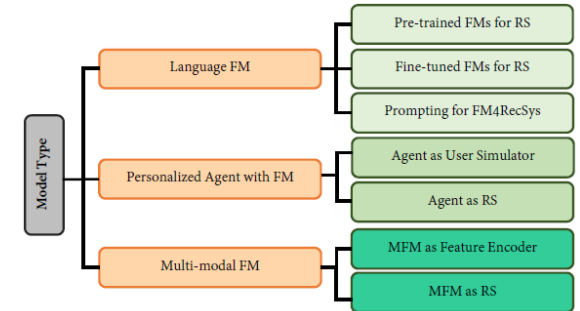


Foundation Models for Recommender Systems: A Survey and New Perspectives

- Personalized Agents with FM
 - Agent as User Simulator
 - Use FMs to simulate user behavior for real-world recommendation
 - LLM-powered agents as human proxies
 - Agent as RecSys
 - “... leveraging the robust capability of FMs, including reasoning, reflection, and tool usage for recommendation.”
 - The FM is used as the “brain” in these architectures
 - Architectures incorporate various tools and memory

Foundation Models for Recommender Systems: A Survey and New Perspectives

- Multi-Modal FMs (MFMs) for RecSys
 - MFMs as Feature Encoder
 - Use representation capability to extract features from text, images, audio
 - MFMs as RecSys
 - Use MFM for multi-modal data processing and recommendation generation
 - e.g., allow multi-modal prompts





A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys)

Yashar Deldjoo
Polytechnic University of Bari, Italy
deldjooy@acm.org

Zhankui He, Julian McAuley
University of California, San Diego,
USA
{zhh004;jmcauley}@ucsd.edu

Anton Korikov, Scott Sanner
University of Toronto, Canada
{anton.korikov;ssanner}@mie.utoronto.ca

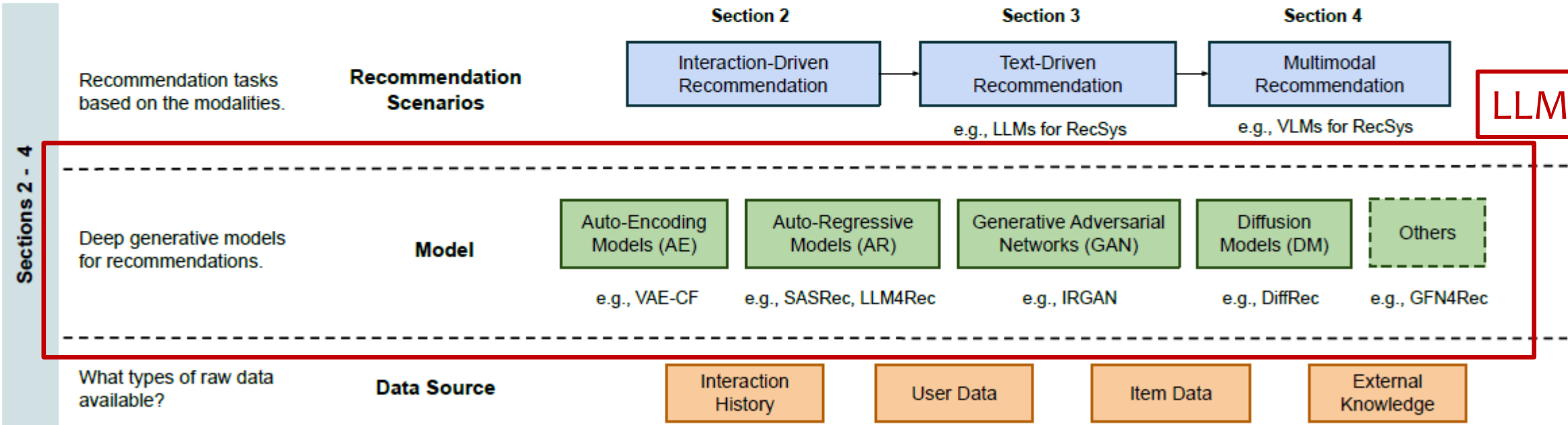
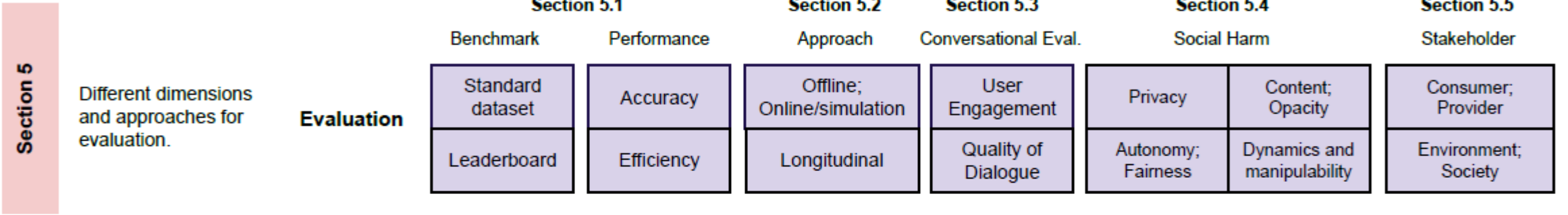
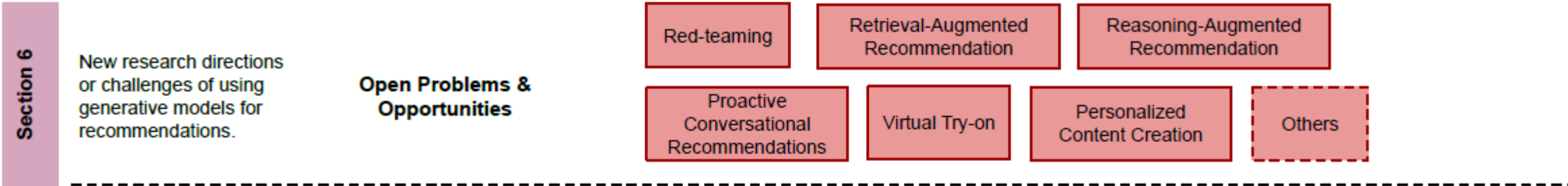
Arnau Ramisa, René Vidal
Amazon, USA
arnaauramisa@gmail.com
vidalr@seas.upenn.edu

Maheswaran Sathiamoorthy
Bespoke Labs
mahesh@smahesh.com

Atoosa Kasirzadeh
University of Edinburgh
atoosa.kasirzadeh@gmail.com

Silvia Milano
University of Exeter and LMU
milano.silvia@gmail.com





A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys)

- Key approaches for LLMs for RS
 - Encoder-only LLM Recommendation
 - Recommendation as Dense (Similarity-based) Retrieval, e.g., using BERT
 - LLM-based Generative Recommendation
 - Prompt-based approaches
 - LLM/RS Hybrids
 - RAG, LLM-based embeddings
 - Conversational Recommendations

A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys)


- Focus on generative Multimodal RS
- Why:
 - Multimodal information (e.g., images) is often available, should be leveraged, e.g., for cold-start problem
 - Novel use cases,
 - e.g., find suitable table given the shape of other furniture
 - e.g., multimodal conversational shopping assistant

Common Challenges – Across surveys

- Inputs, training
 - Prompts can be limited
 - Training efficiency, API costs
- Outputs
 - Biases (popularity, fairness, ...), privacy, explainability
 - Controlling the output (hallucination)
 - Output latency
- Evaluation
 - Datasets like MovieLens might have been seen in input

Part II: Own works

Overview

- 
- LLMs for Sequential Recommendation
 - An **offline** study
 - Explores different paradigms according to the literature surveys
 - ChatGPT as a recommender
 - Focusing on **human evaluation** of a conversational recommender
 - Fashion recommendation
 - No LLMs, but uses image generation
 - Outlook on future **multimodal FMs**

Leveraging Large Language Models for Sequential Recommendation

JESSE HARTE, Delivery Hero Research, Germany and Delft University of Technology, The Netherlands

WOUTER ZORGDRAGER, Delivery Hero Research, Germany

PANOS LOURIDAS, Athens University of Economics & Business, Greece

ASTERIOS KATSIFODIMOS, Delft University of Technology, The Netherlands

DIETMAR JANNACH, University of Klagenfurt, Austria

MARIOS FRAGKOULIS, Delivery Hero Research, Germany

Improving Sequential Recommendations with LLMs

ARTUN BOZ, WOUTER ZORGDRAGER, ZOE KOTTI, and JESSE HARTE, Delivery Hero Research, Germany

PANOS LOURIDAS, Athens University of Economics & Business, Greece

DIETMAR JANNACH, University of Klagenfurt, Austria

MARIOS FRAGKOULIS, Delivery Hero Research, Germany

Sequential Recommendation

- A relevant real-world problem setting
 - Given a sequence of past observed interactions of a user
 - Predict the next item the user will interact with
- Application areas
 - E-commerce, music, videos, point-of-interest recommendation
- Variants
 - Session-based recommendation (of anonymous users)
 - Session-aware recommendation (with knowledge about user IDs)
 - General sequential recommendation (w. user IDs but no sessions)

Sequential Recommendation

Given: a sequence of interactions $S = S_1, S_2, \dots, S_n$ of length n

Predict the next interaction S_{n+1} in the sequence



Algorithmic Approaches

- **History:** Markov Models, Sequential Rules
- **Trending:** Neural Methods
 - Using RNNs, attention, transformer, convolution, deep RL methods
 - Popular:
 - SASRec, BERT4Rec
 - Originally designed for session-based problems:
 - GRU4Rec
- **Relevant baseline:** Nearest-neighbor methods

Leveraging LLMs for Sequential RS

1. **LLMSeqSim :**

- Encode item descriptions via LLM embeddings
- Find semantically similar items in the embedding space

2. **LLMSeqPrompt:**

- A prompt-based approach
- Using a fine-tuned LLM

3. **LLM2Sequential**

- Initialize BERT4Rec or other models with item embeddings with LLM
- Then, train and use the models as before

1) LLMSeqSim

User history

Heinz Ketchup - 500g, Classic tomato flavor, No artificial colors or preservatives, Gluten-free, BPA-free bottle.

Quaker Oats - 1kg, Whole grain rolled oats, High in fiber, Source of protein, Heart-healthy, Resealable bag.

Chobani Greek Yogurt - 170g, Vanilla flavor, 0% fat, No added sugar, Live and active cultures, Non-GMO.

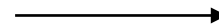
Product catalog

Barilla Spaghetti - 500g, Durum wheat semolina, Non-GMO, Cooks in 10 minutes, Suitable for vegetarians, Product of Italy.

Green Giant Sweet Corn - 340g, Whole kernel, Naturally sweet, No added salt, Non-BPA can, Steam-packed for freshness.

...

Retrieve
and reduce



0.06, 0.76, 0.09, 0.14, 0.51, 0.54, 0.09, 0.37, 0.78, 0.08, 0.35, 0.84

0.57, 0.91, 0.79, 0.44, 0.54, 0.12, 0.29, 0.64, 0.55, 0.92, 0.01, 0.69

0.59, 0.44, 0.41, 0.12, 0.09, 0.55, 0.54, 0.94, 0.94, 0.39, 1.00, 0.58



Aggregate



0.08, 0.42, 0.98, 0.47, 0.29, 0.2, 0.31, 0.36, 0.41, 0.16, 0.8, 0.57



Compare
with each



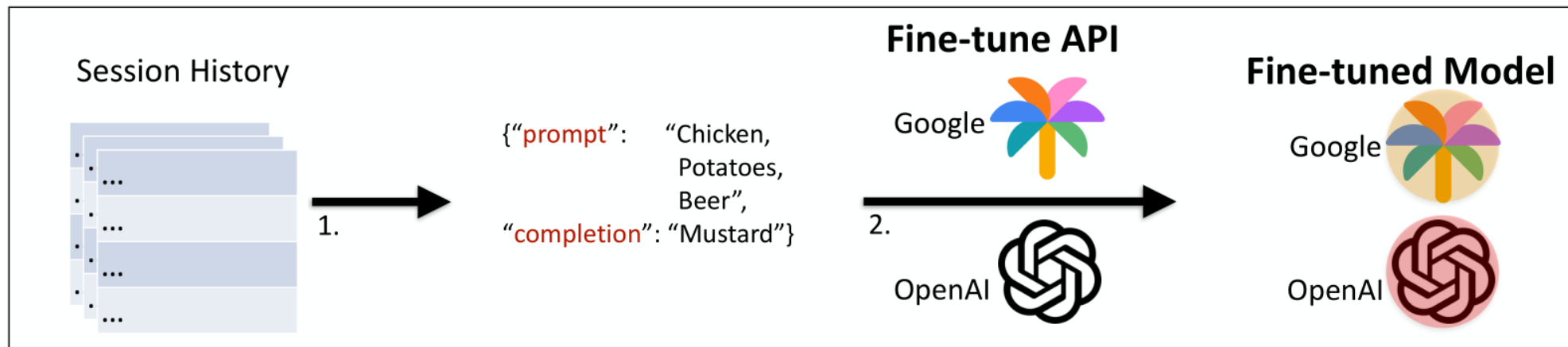
0.23, 0.85, 0.8, 0.61, 0.53, 0.8, 0.61, 0.52, 0.43, 0.95, 0.77, 0.45



0.56, 0.66, 0.25, 0.49, 0.61, 0.11, 0.92, 0.58, 0.35, 0.44, 0.83, 0.36

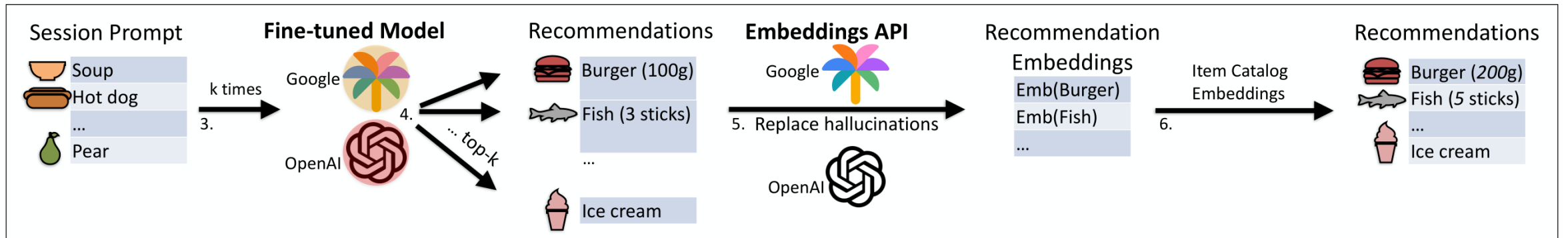
2) LLMSeqPrompt

- **Initial step:** fine-tuning with past sessions
 - Alternative tasks:
 - (1) generate single recommendation, (2) generate list, (3) rank given list



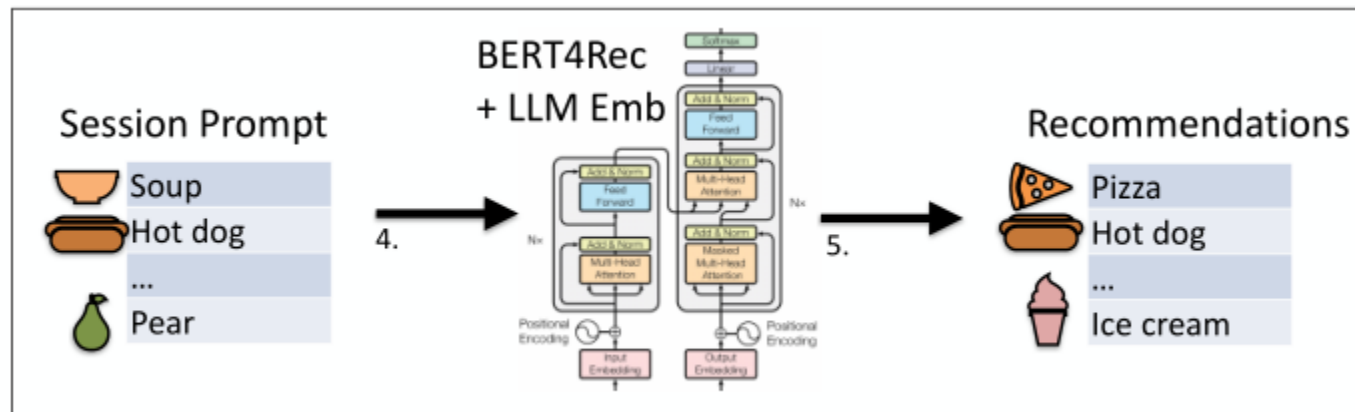
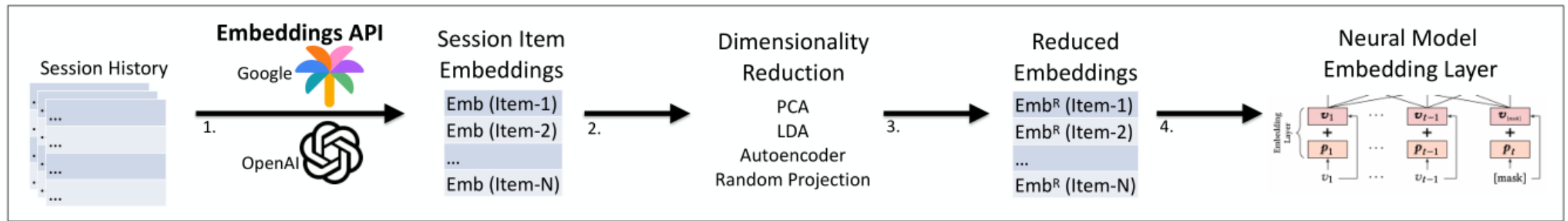
2) LLMSeqPrompt

- **Second step:** Repeatedly prompt fine-tuned LLM
 - Create embeddings from results
 - Replace hallucinations, map back to item catalog



3) LLM2Sequential

- Generalization: **LLM2Sequential** (GRU4Rec, SASRec, BERT4Rec)
 - Initialize item embeddings with LLM embeddings, then recommend



Experiments

- Three sessionized e-commerce datasets
 - Amazon Beauty, Steam, a proprietary from Delivery Hero
 - Up to 250k sessions, up to 38k items
 - Quite different characteristics
- Metrics
 - MRR, NDCG @ 10 and 20
 - Coverage, Popularity

Main outcomes


1. LLM2BERT4Rec is highly effective across datasets
 - best-performing model across datasets
 - e.g., NDCG improved by 15-20% over plain BERT4Rec
2. LLMSeqSim can lead to good results in some cases
 - depends on data, works well for Amazon Beauty, maybe items from the same brand are purchased in one session
 - performs poorly for other dataset like Steam

Summary of Findings

- Main findings
 - Experiments show huge potential of incorporating world knowledge encoded in LLMs
 - Simple prompting approaches (after fine-tuning) may work well in some domains
- Ongoing and future experiments
 - Try different LLMs, but also smaller models (BERT)
 - Explore more datasets
 - Consider multimodal inputs


Overview

- LLMs for Sequential Recommendation
 - An **offline** study
 - Explores different paradigms according to the literature surveys
- • ChatGPT as a recommender
 - Focusing on **human evaluation** of a conversational recommender
- Fashion recommendation
 - No LLMs, but uses image generation
 - Outlook on future **multimodal FMs**



ChatGPT as a Conversational Recommender System: A User-Centric Analysis

AHTSHAM MANZOOR, SAMUEL C. ZIEGLER, KLAUS M. PIRKER GARCIA, and DIETMAR
JANNACH, University of Klagenfurt, Austria



ChatGPT as a Recommender System

Recommend a movie, knowing that I like "Top Gun: Maverick". Be brief.



Check out *Edge of Tomorrow*. It has Tom Cruise, thrilling action, and a strong sense of duty, with a sci-fi twist that keeps the intensity high throughout.



ChatGPT as a Recommender

- **Intuition and observations**

- Even without fine-tuning, ChatGPT makes reasonable recommendations, at least in the movie domain
- ChatGPT can thus be seen to act as a **Conversational Recommender System (CRS)**

Evaluating recent (non-LLM) CRS

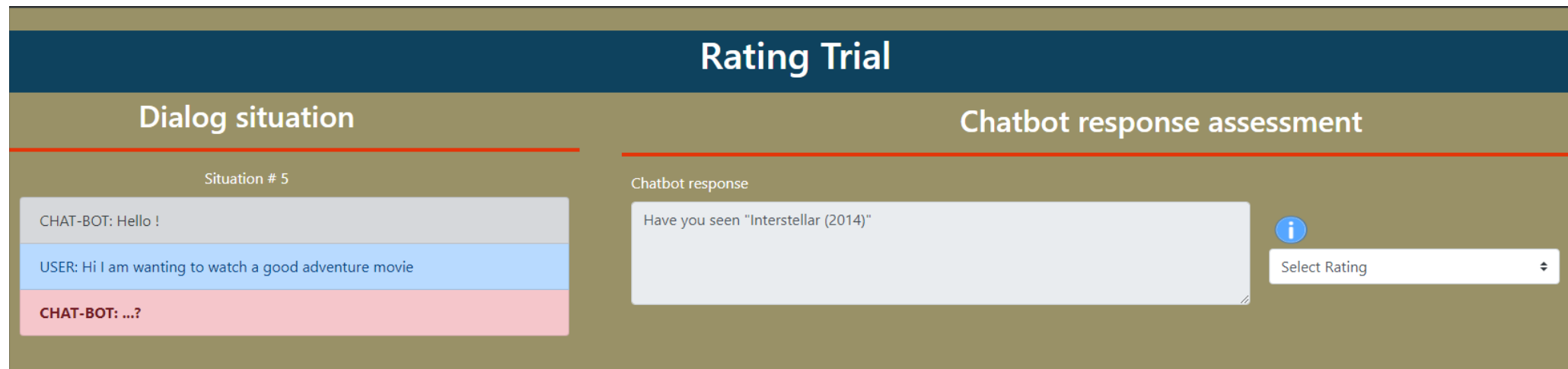
- Evaluating CRS is challenging
 - Linguistic metrics (e.g., BLEU, ROUGE) often not usable
 - Requires humans in the loop
- Own previous research
 - Pre-LLM neural approaches frequently return responses that are not meaningful
 - Up to 30% of utterances not meaningful
 - Such performance numbers can be obtained and improved upon with simpler retrieval-based, heuristic methods

Is ChatGPT a good movie recommender?

- Research question:
 - How well does ChatGPT perform compared to previous academic CRS?
- Methodology:
 - Conduct a **user study**, where participants assess the **meaningfulness of the responses** of a CRS in a given dialog situation
 - Rely on proven research design
 - Comparison between ChatGPT and best previous model
 - Use ReDIAL data set for the evaluation
 - Contains 10,000 dialogs between a movie seeker and a human recommender

User study details

- Rating interface
 - Set of 70 randomly selected dialog situations from ReDIAL
 - Chatbot response
 - Either by ChatGPT (through a prompt) or our CRB-CRS system
 - User task
 - Rate the meaningfulness of the response



The screenshot displays a 'Rating Trial' interface. It is divided into two main sections: 'Dialog situation' on the left and 'Chatbot response assessment' on the right. The 'Dialog situation' section, titled 'Situation # 5', shows a chat history with three messages: 'CHAT-BOT: Hello !', 'USER: Hi I am wanting to watch a good adventure movie', and 'CHAT-BOT: ...?'. The 'Chatbot response assessment' section shows a 'Chatbot response' of 'Have you seen "Interstellar (2014)"'. To the right of the response is an information icon and a 'Select Rating' dropdown menu.

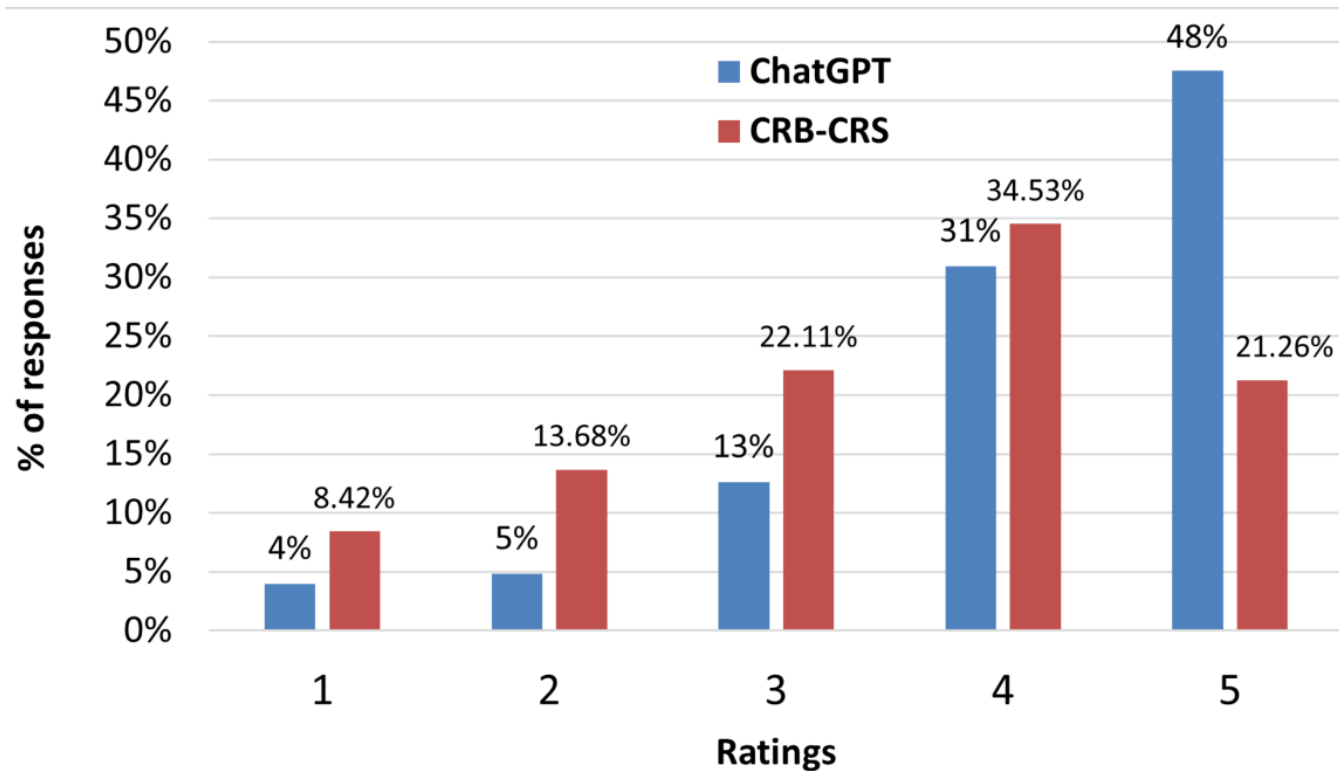
Dialog situation	Chatbot response assessment
<p>Situation # 5</p> <p>CHAT-BOT: Hello !</p> <p>USER: Hi I am wanting to watch a good adventure movie</p> <p>CHAT-BOT: ...?</p>	<p>Chatbot response</p> <p>Have you seen "Interstellar (2014)"</p> <p>Select Rating</p>

Experiment details

- Experiment execution
 - Done by crowdworkers on Amazon Mechanical Turk
 - 100 participants in each group
 - 475 assessments after removing unreliable participants
- Interpretation of “meaningfulness”
 - Room for interpretation for participants
 - Can refer to type of answer (e.g., greeting after a question was asked) or the recommendation itself (e.g., wrong genre)

Results

- Main result
 - ChatGPT largely better than previous best approach



Additional observations


- Looking at details
 - ChatGPT almost never returned meaningless answers
 - In a few cases, in longer utterances that mention several movies, ChatGPT missed the context
- Implications and observations
 - ChatGPT probably difficult to beat with non-LLM approaches for general settings such as movie recommendation
 - ChatGPT is also very good in providing convincing explanations

Outlook and Future Directions

- Outlook
 - ChatGPT (currently) has limitations as CRS, e.g.,
 - No access to real-time catalog of items
 - No long-term user profiles stored
 - Possible hallucination
- Future directions
 - Explore ChatGPT for alternative, more niche domains
 - Explore alternative LLMs
 - Combine with existing RS

Overview

- LLMs for Sequential Recommendation
 - An **offline** study
 - Explores different paradigms according to the literature surveys
- ChatGPT as a recommender
 - Focusing on **human evaluation** of a conversational recommender
- • Fashion recommendation
 - No LLMs, but uses image generation
 - Outlook on future **multimodal FMs**



Fashion Image-to-Image Translation for Complementary Item Retrieval

MATTEO ATTIMONELLI, Politecnico Di Bari, Italy

CLAUDIO POMO, Politecnico Di Bari, Italy

DIETMAR JANNACH, University of Klagenfurt, Austria

TOMMASO DI NOIA, Politecnico Di Bari, Italy



Problem: Complementary Item Retrieval

- **Context:** Fashion Domain
 - General ask of **outfit recommendation**, i.e., recommend a set of matching fashion items
 - Special case: **top-bottom recommendation**,
 - Given a top item (e.g., a t-shirt), recommend a matching bottom item (e.g., pants)
- **Challenge**
 - Compatibility of fashion items may be subjective and context-dependent

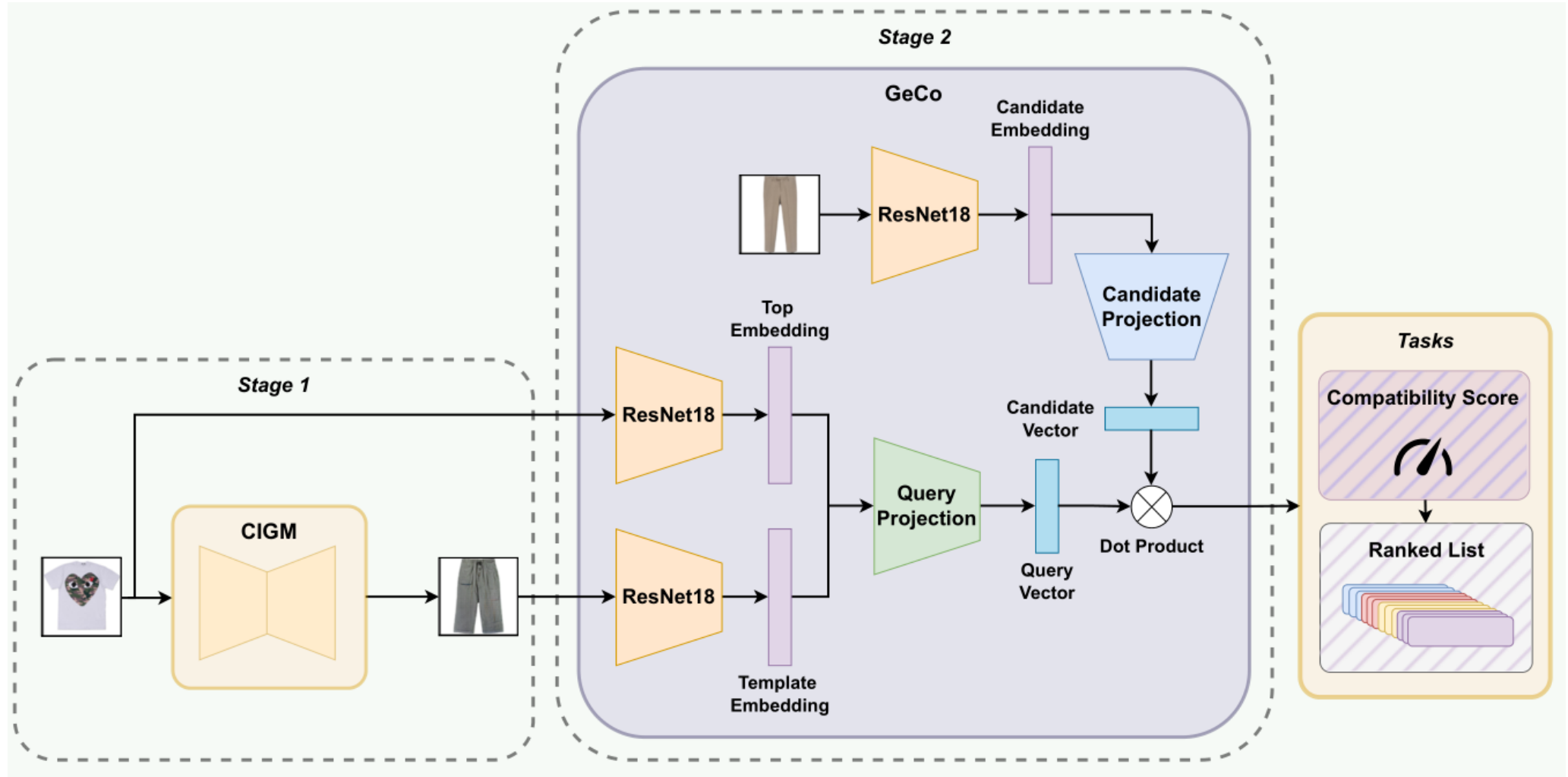
Top-Bottom Retrieval: Literature

- Small set of datasets with matching outfits exist
- Existing approaches
 - Based on the Bayesian Personalized Ranking (BPR) framework
 - Leveraging also visual and textual information
 - Sometimes assumed to lead to modest performance
 - Recently: **Generative** approaches
 - Generate images of suitable bottom items (called “templates”)
 - Leverage the image information for improved compatibility modeling

Generative Approaches: Limitations

- Limitations of recent proposals
 - Images used **solely** for compatibility modeling
 - Quality and realism of generated images not considered for retrieval task
 - Sometimes convergence issues
- Own proposal: **GeCo**
 - Develop a two-stage architecture that leads to high-quality images
 - Use these images as additional input to the retrieval task
 - Ensure scalability as far as possible
 - Challenges of diffusion models

Proposed Architecture: GeCo



Experiments

- Datasets
 - Two existing dataset for top-bottom retrieval
 - Contributed a new one (subset of Taobao dataset), larger than previous top-bottom datasets
 - 30k tops, 25k bottoms, 76k positive pairs
- Baselines
 - BPR-DAE: A content-based clothing matching scheme
 - MGCM: A GAN-based compatibility modeling scheme
 - Pix2PixCM: Like MGCM but with a Pix2Pix model

GeCo – Realism of Generated Templates



(a) Conditioning tops.



(b) Ground-truth bottoms.



(c) Generated bottoms with the proposed Generative Model.

Results

- The proposed model leads to good performance
 - In terms of AUC and MRR
- The older BPR-DAE model surprisingly often second-best

Model	FashionVC		ExpReduced		FashionTaobaoTB	
	<i>AUC</i>	<i>MRR</i>	<i>AUC</i>	<i>MRR</i>	<i>AUC</i>	<i>MRR</i>
GECo	0.7621	0.0307	0.7295	0.0151	0.8491	0.0126
MGCM	0.5939	0.0098	0.4674	0.0075	0.7508	<u>0.0095</u>
Pix2PixCM	0.5000	0.0080	0.5136	0.0078	0.6934	0.0056
BPR-DAE	<u>0.6590</u>	<u>0.0200</u>	<u>0.6457</u>	<u>0.0130</u>	<u>0.7531</u>	0.0052
RANDOM	0.4908	0.0081	0.4908	0.0063	0.5046	0.0014

Conclusion

- Generative models can help improve recommendation performance
- In our approach:
 - Generated images are input to the second-stage compatibility modeling and ranking process
- Future work:
 - Incorporation of textual information as input
 - Personalization

Summary and Outlook

Summary

- Huge opportunities exist for the development of RS with LLMs
- We reviewed existing categorizations of LLM-enhanced recommendations
- We discussed recent research efforts
 - Sequential recommendation and offline evaluation
 - Conversational recommendation and human evaluation
 - Multimodal recommendation

Future Directions: Multimodal RS

- As identified as in at least two surveys



Recommend suitable pants



Using Google Gemini



Recommend suitable pants



The smiley face t-shirt you sent leans towards the casual side, so you'll want to pair it with pants that are also casual. Here are a few options:

- **Jeans:** A classic and versatile option that goes well with almost anything. You can dress them up or down depending on the occasion. For a more casual look, try a pair of relaxed-fit or boyfriend jeans.



relaxed fit jeans

Challenges and Opportunities: Evaluation

- Various challenges identified in existing surveys
- Key pain point in academia: **Evaluation**
 - Almost all published work based on offline experiments
 - No clear correspondence between offline accuracy and value for consumers and business
 - Furthermore, a huge fraction of research works is flawed, even at top-level conferences
- LLM-based RecSys as an **opportunity for change**
 - Put the human and other stakeholders in the center
 - Understand the value of available models first before creating new ones

-
- Thank you for your attention
 - Enjoy the conference!

 - Contact:
 - dietmar.jannach@aau.at

 - Slides:
 - <https://tinyurl.com/kdd24-genai>

